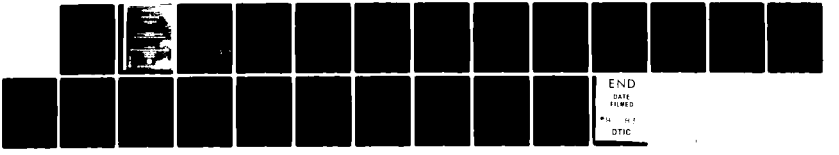
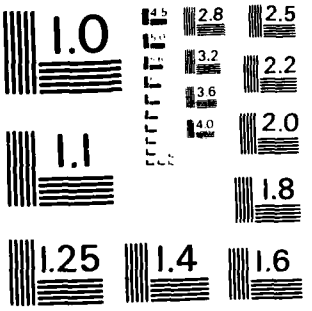
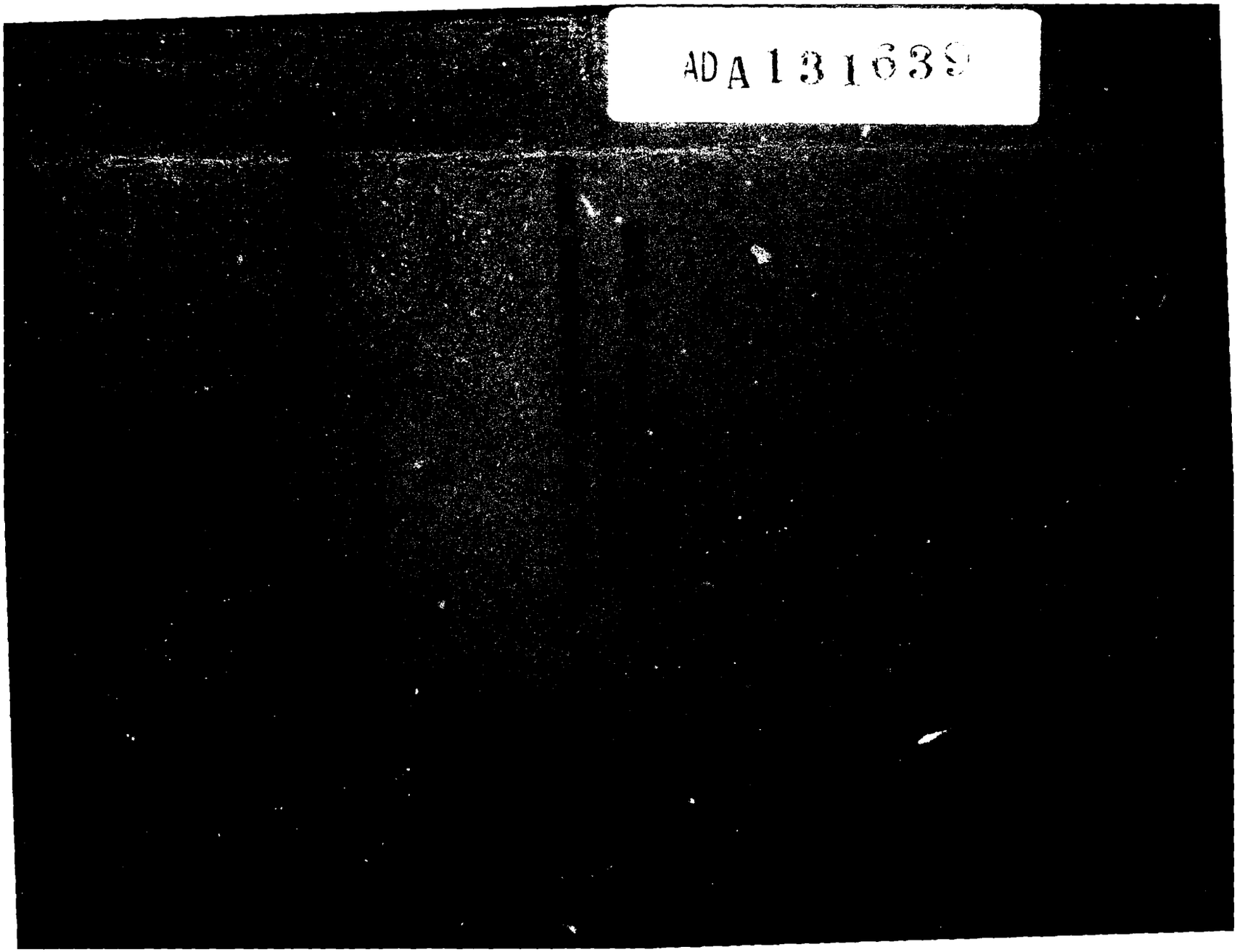
AD-A131 639 SEARCH AMONG QUEUES(U) STANFORD UNIV CA CENTER FOR RESEARCH ON ORGANIZATIONAL EFFICIENCY A GLAZER ET AL. JUN 83 TR-406 N00014-79-C-0685 1/1

UNCLASSIFIED F/G 12/1 NL



END
DATE FILMED
DTIC





SEARCH AMONG QUEUES

by

Amihai Glazer and Refael Hassin

Technical Report No. 406
June, 1983

Prepared Under
OFFICE OF NAVAL RESEARCH GRANT ONR-N00014-79-C-0685

THE ECONOMIC SERIES

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES
Fourth Floor, Encina Hall
Stanford University
Stanford, California
94305

SEARCH AMONG QUEUES*

by

Amihai Glazer**and Refael Hassin***

1. Introduction

Customers must often wait to obtain some service or good, the wait usually being longer the greater the total number of persons who have yet to be served. This phenomenon is especially common in transportation markets, such as the trucking, household moving, and bus industries. Nor is it unknown for customers to have to wait for the services of accountants, lawyers, dentists, phsicians, and -- most importantly -- plumbers. When for some reason shortages occur, rationing is often accomplished by queuing; new models of automobiles and computers have lately been subject to such shortages.

Recent research has shown that explanations of market behavior should incorporate customer's aversion to such waiting (see DeVany [1976], DeVany and Saving [1977], Koenigsberg [1980]). The fundamental idea is that a customer who arrives at some facility and finds a long queue may find it worthwhile to balk, incur some search cost, and find a shorter queue at some other facility. This in turn means that excess capacity may result not from oligopoly or inefficiencies, but from a socially beneficial effort to reduce consumers' waiting costs.

---

* This research was supported by the office of Naval Research Contract ONR-N000[illegible]-79-C-0685 at the Center for Research on Organizational Efficiency, Institute of Mathematical Studies in the Social Sciences, Stanford University, Stanford, California.

**School of Social Science, University of California at Irvine, California, 92717.

***Statistics Department, Tel-Aviv University, Tel-Aviv, 69978, Israel.

Models of such markets must incorporate some notion of consumer's search behavior, i.e., the conditions under which a customer decides to leave one queue in search of another. Two extreme cases will be studied here. In one case we suppose there is an indefinitely large number of facilities, so that a customer who discovers the length of one queue learns nothing new about the length of any other queue. In the other case, we assume the existence of only two facilities, so that a customer who finds a long queue at one of them would rationally expect the other queue to be long as well.

We are especially interested in determining whether optimal search possesses the reservation length property: that is, will a customer engage in further search if and only if he finds a queue length greater than some critical value? This property is analogous to the "reservation wage" property found in studies of job search -- certainly the most elegant and useful result commonly accepted in that literature. Indeed, in the absence of some such property (and in our case it may be absent), the analysis of models incorporating search becomes very difficult, and virtually forecloses the possibility of finding general analytic solutions.

The questions we ask have been addressed in some form or another in earlier papers. Koenigsberg [1966] studied jockeying in queues, assuming that customers incur no search costs. The same author [1980] also studied the characteristics of a duopolistic market; but rather than explicity analyzing consumers' decisions, Koenigsberg made the simple and useful assumption that a customer is more likely to engage in search the longer the queue in which he finds himself. Essentially the same

assumptions were made by Kornai and Weibull [1980]. DeVany [1976], in a pathbreaking article, examined a duopoly in which customers engage in search; to make the problem tractable he assumed that the queue length at one of the firms is not correlated with the queue length at the other. In a later article modeling competition between trucking firms, DeVany and Saving [1977] implicitly assumed that a customer can balk only once, and found that the competitive equilibrium level of excess capacity is identical to the socially optimal level.

We are not the first, of course, to inquire whether optimal search always possesses a property like the reservation wage one. Aharon and Veendorp [1983] showed that a customer who faces a budget constraint will stop searching only if the price he discovers is greater than some critical level; this critical level, however, will be a function of the number of shops the customer had already visited. Rothschild [1974] examined a search model in which customers are initially uncertain about the distribution of prices on the market, and learn about that probability distribution as they search from it. He shows that even then, in many instances, optimal search will follow the reservation price strategy. Our approach is related to, but by no means identical to, his. We suppose that customers do know the probability distribution of queue lengths, but that the queue lengths at the different facilities need not be independent; moreover, we let this distribution depend on the search strategies customers adopt.

Our assumptions are set forth in Section 2. Section 3 concerns search when the number of facilities is arbitrarily large and shows that

the reservation length property (RLP) does hold then. Section 4 examines the more difficult problem of search among two queues. In this case the optimal search strategy is not always RLP. Section 5 offers a short conclusion.

2. Assumptions

Suppose that in a market there exist several facilities. The queue length at each such facility will vary over time in a random manner as new customers join the queue and the customer (if any) who is at the head of the line has his service completed and leaves the facility. One can usefully think of any one queue in terms of a stack of trays, where every so often a tray can be added to the bottom or removed from the top.

Most reasonably, customers at any facility are served in a First-Come-First-Served order, so that any one customer's expected waiting time is proportional to the number of customers ahead of him. Suppose also that a customer who has joined some queue can discover at zero cost the length of the queue there; to determine the length of any other queue, however, he must incur some search cost. A search model must determine how a customer should decide whether or not to engage in search, given the queue length at the facilities he had already visited. A general search model would have to be a most complicated one, and, for example, may incorporate the following features.

1) The lengths of any two queues need not be <u>independently</u> distributed random variables; instead, one would expect that if one facility is heavily congested and if some consumers engage in search, the other facilities will be congested as well.

2) Moving from one queue to another might require some time, and during that time the queue lengths at the facilities will change. This means that a customer would have to choose among three courses of action: a) staying where he is; b) returning to re-check a queue he had already visited; c) checking the length of a queue he had never visited.

3) Customers may decide to balk not only upon fist arriving at some facility, but also after having spent some time waiting in line. A customer's decision can then depend not only on the length of his queue at any instant, but also on the length of time that has elapsed since he had checked some other queues.

4) A customer may have several ways of conducting search. He can move to another queue without obtaining any prior information, or he might be able to telephone in advance to learn about the queue length.

Solving a model which incorporates all these features appears to be a herculean task. But for our purposes it is not necessary to solve the most general model possible. Instead, we wish to describe a simple model which captures the essential elements involved in queues, and which allows us to determine whether optimal search has the reservation length property. If the answer to this question is in the negative for the simple model, it must also be so for a more general model.

We assume that there are m facilities, each with a single server. The service time is exponentially distributed with mean $1/\mu$. The stream of new customer arrivals to the system can be described by a Poisson distribution with parameter $m\lambda$. All customers who enter the system are eventually served; to ensure that this is feasible we require

that $\mu > \lambda$. We assume that these parameters, as well as the queue length and the waiting time distributions, are known to all customers. Customers are served in a First-Come-First-Served order. A customer arriving at some facility immediately discerns the length of the queue there. He can then join the queue, or else search for another one. A customer cannot leave a particular queue after having decided to join it.

The cost of search is c. After spending that amount a customer can determine the length of some queue other than the one at which he had just arrived, and then instantaneously and without further cost joint the shorter of the two queues under consideration. Multiple searches are permitted.

A customer's waiting cost has a constant value of 1 per unit of time. All customers are identical, risk-neutral, expected utility maximizers. Each wishes to minimize the sum of his expected waiting and search costs.

3. Infinite Number of Servers

This section considers a market consisting of an indefinitely large number of service facilities, and demonstrates that under such conditions an individual's optimal search strategy possesses the reservation length property. We first determine the steady state properties of a queuing system in which customers adopt such a strategy, then determine an individual's optimal strategy given these properties of the system, and finally show that RLP search yields a consistent solution.

Suppose that a customer balks whenever he encounters a queue of length greater than or equal to B. That is, a customer will join a particular queue only if its length is B - 1 or less when he visits it; otherwise he will continue searching. This assumption implies that no queue will be longer than B. Define $\alpha = \alpha(B)$ as the steady state propoertion of queues at any instant with a length of exactly B; $(1 - \alpha)$ of the queues are shorter than B. The length of the queue at any facility will, of course, vary over time. The rate per facility at which customers appear in the system is λ. As customers will only join queues shorter than B, the total average rate of customer arrivals at a queue with less than B customers is $\lambda + \lambda\alpha/(1 - \alpha) = \lambda/(1 - \alpha)$.

We have in essence described a queuing system with balking, where each queue has a waiting room of capacity B. The steady-state behavior of such a system is well known (see e.g., Naor [1969]). Let $p_1,\ldots,p_B$ be the steady-state probabilities of the length of the queue at any facility, and define $\rho \equiv \lambda/(1 - \alpha)\mu$. Then $p_i = \rho^i p_0$ for $i = 1,\ldots,B$, $p_0 = (1 - \rho)/(1 - \rho^{B+1})$, and $p_B \equiv \alpha = (1 - \rho)\rho^B/(1 - \rho^{B+1})$. By the central limit theorem, for a sufficiently large number of facilities in the system the actual propoertion of queues with length i at any instant differs from the steady-state probabilities p_i by an arbitrarily small amount. This in turn means that if the number of queues is sufficiently large, the lengths of any two queues can be treated as <u>independent</u> random variables.

Given such a queuing system, any one customer's decision is quite simple. A customer who joins a queue with i persons ahead of him will wait an average of v units of time until his service starts. The benefit of

conducting search is proportional to the difference between the lengths of the two queues if a shorter queue is found, and is zero otherwise. Since the lengths of any two queues are uncorrelated, the value of search must be a decreasing function of i: a customer will continue the search if and only if i is larger than some critical value. (The proof is analogous to that used in the job search literature. See Lippman and McCall [1976], pp. 157-163.)

This critical value, called here B, can be found by using the attributes of the queuing system described above. Let $F(B)$ be a customer's expected sum of waiting and search costs, when his policy is to join a queue if and only if its length is less than B. Then,

$$F(B) = \sum_{i=0}^{B-1} (\frac{i}{\mu}) p_i + \alpha(a + F(B)) \quad ,$$

so that

$$F(B) = \frac{1}{1 - \alpha} \sum_{i=0}^{B-1} (\frac{i}{\mu}) p_i + \alpha c \quad .$$

The reservation queue length is that value of B which satisfies the condition that $(B - 1)/\mu \leq c + F(B) < B/\mu$.

The analysis of markets with a very large number of facilities is therefore quite simple. One can assume that customers will balk at any queue whose length is greater than some critical level. Indeed, this property holds even if the inter-arrival time has a distribution more general than the exponential. It is important, however, that the service rate be exponential; if it is not then a customer's optimal strategy depends in part on the residual service time and search will not be RLP.

Our model can be interpreted as referring not only to search among many facilities, but also to a customer's decision of whether to balk and to return to the facility later. Suppose there exists only one service facility and that some customer arrives there to find a long queue. He can either immediately join the queue, or else he can balk, and at a cost of c dollars return later. If this customer returns at some arbitrarily chosen time far in the future, the queue length t] will be uncorrelated with the queue length at the time he balks. The ıracteristics of his search decision are then identical to those described : . ; section, and his decision of whether or not to balk will have the reservation length property.

We should also note that search, either over time or over facilities, involves a positive externality. A customer who searches obviously expects to reduce his expected costs. But in the course of search he may find an empty facility, and may have his service finished before any other customer comes to that queue. This would reduce the expected waiting time of future customers, a benefit which the potential searcher does not include in his calculus.

4. A Market with Two Queues

Suppose that the assumptions made above hold, but that there exist only two service facilities. In that case, the lengths of the queues are not independent random variables: if the queue is long at one facility, it is probabily long at the other one as well. This means that, in distinction to the case discussed above, the benefits of search need not

be an increasing function of the queue length at a customer's current facility, and that the optimal search strategy need not be RLP.

We shall prove this assertion in several steps. First, we briefly describe the rules of the queuing system. Second, for any given search strategy that is used by all customers, we find the probability distributions of the queue lengths at the two facilities. Third, we give some numerical solutions of these equations for the particular case in which each customer searches only if the queue he encounters is longer than some critical value. Fourth, we use these solutions to determine a customer's expected utility from using several specified search strategies, and find that the optimal search strategy will not always be RLP.

A customer who wishes to be served will first visit one of the two facilities; suppose that he is equally likely to first visit each of the two facilities. He then observes the length of the queue there, i. At this point he must decide whether or not to search. If he does not search, he joins the queue and incurs an expected waiting cost of v. By spending an amount c, he can determine the length, j, of the other queue, and then join the shorter queue; his expected wait is therefore $\min [i,j]/\mu$.

A customer's search strategy can thus be defined by specifying whether he does or does not search when at the first facility he visits he finds a queue of length i. Let $S_i = 1$ mean that he does search when that length is i, and let $S_i = 0$ mean that he does not then search. The behavior of the queuing system is a function of this search strategy. The length of a queue can change in four ways: 1) its length decreases by

one when a customer who is being served has his service completed; 2) its length increases by one if a new customer joins that queue and did not conduct search; 3) its length increases by one if a new customer visits the facility, conducts search, and finds the first queue to be shorter, and 4) its length increases by one if a customer first visited the other facility, conducted search, and decided to move to this queue.

Let μ be the service rate at each facility, let 2λ be the arrival rate of new customers to the system, let dt be an infinitesimal increment of time, let values of S_i $(i = 0,1,2,\ldots)$ define a customer's search strategy, and let $p_{ij}(t)$ be the probability that at time t, i customers are at the first facility, and j customers are at the second. The following equations describe the transitions of these probabilities:

$$p_{ij}(t + dt) = [1 - (L_3\mu + \lambda)dt]p_{ij}(t) + L_1 dt p_{i,j-1}(t) + L_2 dt p_{i-1,j}(t) + \mu dt[p_{i,j+1}(t) + p_{i+1,j}(t)] \quad (1)$$

$$\text{for } i = 0, 1, 2, \ldots$$

$$j = 0, 1, 2, \ldots$$

where

$$L_1 \equiv \begin{cases} \lambda & \text{if } i = j - 1 \\ \lambda(1 - S_{j-1}) & \text{if } i < j - 1 \\ \lambda(1 + S_i) & \text{if } i > j - 1 \end{cases}$$

$$L_2 \equiv \begin{cases} \lambda & \text{if } i = j - 1 \\ \lambda(1 - S_{i-1}) & \text{if } i - 1 > j \\ \lambda(1 + S_j) & \text{if } i - 1 < j \end{cases}$$

$$L_3 \equiv \begin{cases} 2 & \text{if } i \neq 0 \text{ and } j \neq 0 \\ 1 & \text{if } i = 0 \text{ or } j = 0 \text{ but not both} \\ 0 & \text{if } i = j = 0 \end{cases}$$

and $\quad p_{ij}(t) \equiv 0 \quad \text{if} \quad i = -1 \quad \text{or} \quad j = -1 \quad .$

The steady-state behavior of the system is given by the values of p_{ij} which satisfy the condition $p_{ij}(t) = p_{ij}(t + dt)$ for all i and j.

We have no way of finding explicit solutions of these equations, although we can solve them numerically. Fortunately, for our purposes such solutions will suffice. Recall that our goal is to determine whether the reservation length property holds. We can solve equations (1) for cases in which $S_i = 0$ for $i < B$, and $S_i = 1$ for $i \geq B$; this specification reflects a search strategy that is RLP. Given the steady-state solutions, for any given value of i we can determine a customer's benefit from search. Let $V(i) \equiv [(i/\mu) \sum_{j \geq i} p_{i,j} + (j/\mu) \sum_{j<i} p_{ij}] / \sum_j p_{ij}$. The benefit of search is then $V(i) - c$. Table 1 presents some values of $V(i)$ for given values of B, where $\lambda = 0.5$ and $\mu = 3$.

A necessary condition for the optimal search strategy to be RLP is that $V(i) < c$ for all $i < B$, but that $V(i) > c$ for all $i \geq B$. Consider, however, the values of $V(i)$ given for $B = 6$ in Table 1:

i \ B	1	2	3	4	5	6	7
1	.746	0.820	0.831	0.833	0.833	0.833	0.833
2	1.200	1.676	1.788	1.803	1.805	1.805	1.805
3	1.408	1.401	2.679	2.787	2.799	2.801	2.801
4	1.495	1.493	1.493	3.689	3.788	3.799	3.800
5	1.546	1.536	1.535	1.535	4.693	4.789	4.799
6	1.590	1.556	1.555	1.555	1.555	5.695	5.789
7	1.653	1.573	1.565	1.564	1.564	1.564	6.696
8	1.729	1.586	1.572	1.569	1.569	1.568	1.568

Table 1

$V(7) = 1.564$, $V(6) = 5.695$, and $V(5) = 4.789$. If search is to be conducted whenever $i \geq 6$ but not when $i < 6$, c must in this case be both greater than 4.789 (so that a customer will not search when $i = 5$) and less than 1.564 (so that a customer will search when $i = 7$). As this is impossible, we must conclude that if $\mu = 3$ and $\lambda = 0.5$, there exist no values of c for which the optimal search strategy has the reservation length property with a critical value of 6. Further examination of Table 1, and extrapolation therefrom, shows that if $\mu = 3$ and $\lambda = 0.5$, the optimal search strategy can never have the reservation length property for values of c greater than 1.401. The same negative result can hold for other values of μ and λ.

The intuitive explanation for this is straightforward. Suppose that $B = 6$ and that a customer finds himself at a queue with 6 customers already there. By assumption, the sixth person in the queue had examined

the length of both queues, found the one at the given facility shorter, and therefore joined it. The fact that there are B or more persons in some queue indicates that there are probably many persons at the other facility as well. A rational customer should take this information into account, and may therefore find it worthwhile to search when encountering a queue of length B - 1, but not when finding one of length B.

This does not mean, however, that search can never be RLP. An examination of Table 1 shows that if $c < 0.746$, optimal search is RLP with a critical value of 1. If c lies between .820 and 1.401, a customer will engage in search whenever $i \geq 2$. For values of c lying in the interval (.746, .820), further numerical calculations show that if some customers search whenever $i \geq 1$, and all others when $i \geq 2$, no customer can do better than to adopt either of these strategies. That is, search can take the form of a mixed strategy.

Several features are noteworthy about these solutions.

a) The argument does not really require the assumptions that the customers' arrival process be Poisson. The results are far more general.

b) A customer's decision of when to search does not yield a solution that would minimize the expected costs of all customers. For search involves an externality. If some customer searches for a shorter queue, it becomes less likely that one server is idel while the other queue has a positive length. This in turn decreases the expected number of people in the system and decreases the expected waiting time of other customers. This benefit, however, accrues not to the searcher, but to customers who arrive after him. Similarly, a customer who searches and

joins the shorter of two queues conveys information to later consumers by the fact of his presence at one queue rather than the other. This benefit to future customers is not part of an individual's calculus.

c) The type of argument we used for the case in which there are two queues also applies for a system in which there are a finite number, m, of queues. In any such system, customers who use a reservation length strategy convey information to future customers. Intuitively, then, a new customer who finds a long queue can reasonably suppose that the expected queue lengths at all other facilities are larger than their average. This might induce him to avoid engaging in search. Only if the number of queues is indefinitely large does the central limit theorem ensure that the queue lengths at any two facilities are not correlated.

5. Conclusion

We have been able to determine the optimal search policy among queues if the number of queues is very large. The reservation length property makes the analysis of markets with such queues both elegant and tractable. Such is not the case with a small number of queues. Indeed, we suspect that no useful analytic solutions can be found for the optimal search policies in such cases, and that future research will have to rely on numerical solutions. Yet even in the absence of explicit solutions we did reach one important conclusion: customers have insufficient incentives to search, and even if no firm exercises its market power the free market solution will not be a socially optimal one.

References

Aharon, R. and E.C.H. Veendorp [1983], "Sequential Search with a Budget Constraint," Economic Letters, Vol. 11, No. 1-2, pp. 81-85.

DeVany, A. [1976], "Uncertainty, Waiting Time, and Capacity Utilization: A Stochastic Theory of Product Quality," Journal of Political Economy, June, pp. 523-541.

DeVany, A. and T.R. Saving [1977], "Product Quality, Uncertainty, and Regulation: The Trucking Industry," American Economic Review, September, pp. 583-594.

Koenigsberg, E. [1966], "On Jockeying in Queues," Management Science, January, pp. 412-436.

Koenigsberg, E. [1980], "Uncertainty, Capacity, and Market Share in Oligopoly: A Stochastic Theory of Product Quality," Journal of Business, pp. 151-164.

Kornai, J. and J.W. Weibull [1980], "Mathematical Appendix A: Queuing on the Market," in J. Kornai, Economics of Shortage, Vol. B, Amsterdam: North Holland Publishing Company.

Lippman, S.A. and J.J. McCall [1976], "The Economics of Job Search: A Survey," Economic Inquiry, June, pp. 155-189.

Naor, P. [1969], "On the Regulation of Queue Size by Levying Tolls," Econometrica, pp. 15-24.

Rothschild, M. [1974], "Searching for the Lowest Price when the Distribution of Prices is Unknown," Journal of Political Economy, July/August. Pp. 689-711.

REPORTS IN THIS SERIES

REPORTS IN THIS SERIES

Reports in this Series

376. "Necessary and Sufficient Conditions for Single-Peakedness Along a Linearly Ordered Set of Policy Alternatives" by P.J. Coughlin and M.J. Hinich.

377. "The Role of Reputation in a Regeated Agency Problem Involving Information Transmission" by W. P. Rogerson.

378. "Unemployment Equilibrium with Stochastic Rationing of Supplies" by Ho-mou Wu.

379. "Optimal Price and Income Regulation Under Uncertainty in the Model with One Producer" by M. I. Taksar.

380. "On the NTU Value" by Robert J. Aumann.

381. "Best Invariant Estimation of a Direction Parameter with Application to Linear Functional Relationships and Factor Analysis" by T. W. Anderson, C. Stein and A. Zaman.

382. "Informational Equilibrium" by Robert Kast.

383. "Cooperative Oligopoly Equilibrium" by Mordecai Kurz.

384. "Reputation and Product Quality" by William P. Rogerson.

385. "Auditing: Perspectives from Multiperson Decision Theory" By Robert Wilson.

386. "Capacity Pricing" by Oren, Smith and Wilson.

387. "Consequentialism and Rationality in Dynamic Choice Under Uncertainty" by P.J. Hammond.

388. "The Structure of Wage Contracts in Repeated Agency Models" by W. P. Rogerson.

389. "1980 Abraham Wald Memorial Lectures, Estimating Linear Statistical Relationships by T.W. Anderson.

390. "Aggregates, Activities and Overheads" by W.M. Gorman.

391. "Double Auctions" by Robert Wilson.

392. "Efficiency and Fairness in the Design of Bilateral Contracts" by S. Honkapohja.

393. "Diagonality of Cost Allocation Prices" by L.J. Mirman and A. Neyman

394. "General Asset Markets, Private Capital Formation, and the Existence of Temporary Walrasian Equilibrium" by P.J. Hammond

395. "Asymptotic Normality of the Censored and Truncated Least Absolute Deviations Estimators" by J.L. Powell

396. "Dominance-Solvability and Cournot Stability" by Herve Moulin

397. "Managerial Incentives, Investment and Aggregate Implications" by B. Holmstrom and L. Weiss

Reports in this Series

398 "Generalizations of the Censored and Truncated Least Absolute Deviations Estimators" by J.L. Powell.

399. "Behavior Under Uncertainty and its Implications for Policy" by K.J. Arrow.

400. "Third-Order Efficiency of the Extended Maximum Likelihood Estimators in a Simultaneous Equation System" by K. Takeuchi and K. Morimune.

401. "Short-Run Analysis of Fiscal Policy in a Simple Perfect Foresight Model" by K. Judd.

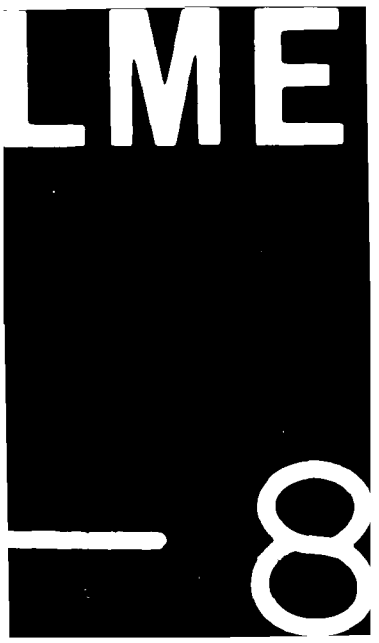
402. "Estimation of Failure Rate From A Complete Record of Failures and a Partial Record of Non-Failures" by K. Suzuki.

403. "Applications of Semi-Regenerative Theory to Computations of Stationary Distributions of Markov Chains" by W.K. Grassmann and M.I. Taksar.

404. "On the Optimality of Individual Behavior in First Come Last Served Queues With Preemption and Balking" by Refael Hassin.

405. "Entry with Exit: An Extensive Form Treatment of Predation with Financial Constraints" by J.P. Benoit.

406. "Search Among Queues" by A. Glazer and R. Hassin


LME
8